Weekly Report

2018.6.4 — 2018.6.10

Job Summary

1. I corrected the chapter arrangement of my graduation thesis according to graduation reply and submit the final version.
2. I read the chapter 3 of *data visualization*, the PPT of *data* and other relevant documentation for the explanation for Singapore students on Wednesday.
3. I designed the initial sketch for the bank transaction display. There are two kinds of design. The two designs both will be completed to choose a better one.
    a) The dynamic effect diagram use a pie chart and the static diagram use a stacked area graph.
    b) Just use a stacked area graph and it can move toward left to display changed data.

4. I read the review of VAST and make a summary. We make a plan for the revise.

Work Plan

1. Complete the preliminary code for bank transaction display and help the Singapore student to finish a part.
2. Define the task and revise the pater.
    a) Sampling Algorithm

i. It applies in the 2D-projection space and we should find some paper to support our opinion that tsne do some helpful work to make the sampling more effective because it makes the vectors close in the high dimension closer in the 2D space.

ii. Compare the different sampling methods.

b) Case Study: It should verify the advantage of the sampling based on the word embedding view.

c) Design: It should support our sampling mehods.

3. Read the paper for report - *Timelines Revisited: A Design Space and Considerations for Expressive Storytelling*

Future work

1. Improve code of band project and make the component.
2. Prepare the report for the week seminar.

Appendix

Summary

Thus the identified strengths of the paper are:

+ basic idea of using Word2Vec model to represent OD flows very interesting (R2, R4) and novel (R3, R4)
+ sampling methods using adaptive blue noise sampling looks useful (R2) and successfully reduce visual clutter (R3)
+ appreciate the method that arranges massive OD flows into 2D space, which makes us understand the distribution and semantic of OD flows easier (R1)
+ means to at least qualitatively evaluate effects of sampling (R4)
+ overall paper structure is sound (R3, R4)

（证明算法的优势-case）

- failure to demonstrate the advantage of the sampling method based on the word embedding view, compared to the spatial distribution or community distribution based sampling (R1, R3), effects of word2vec (R3)

- abstraction process may distort the original geographic shape of the flows, implications of this need further investigation (R3, R4)

- possible introduction of sampling artifacts due to 2D projection of the feature space (R3, R4)

- case studies aren't convincing/fail to illustrate the claimed contributions (R1, R2, R4) and performance of the core NLP approach in corner cases (R3, R4)

- several design choices/selections not explored, in particular, selection of the specific word2vec model (R3)

- "bag of techniques" approach wrt. possible sampling strategies puts the burden of finding the best combination on the user with no help from the system, paper does not provide corresponding guidelines (R3, R4)

- several minor problems with clarity of explanations and/or ambiguous definitions, as pointed out in the respective detail reviews (R1, R2, R3, R4)
- missing references to related works, see respective detail reviews (R2, R4)
- supplemental movie needs to be improved by voice-over or subtitles (R2, R3, R4)
- quality of writing/manuscript formatting needs many improvements (R1, R2, R3, R4)

粉色-算法　　黄色-标注与解释　　绿色-设计　　蓝色-case

1.（3）

- fail to demonstrate the advantage of the sampling method based on the word embedding view, comparing to the spatial distribution or community distribution based sampling.

Major problem

（语义结果近似空间分布，证明基于语义采样的优越性）

- I have concerns about the sampling method. The major assumption of this research is the sampling method based on the word embedding projection can retain the semantic interactions of OD flows. In my opinion, however, the result of Word2Vec model is very similar to community detection method (as we see in Figure 4). For this reason, the result of sampling method based on word embedding view should be very similar to the sampling method based on spatial distribution or community distribution. Please demonstrate the advantage of the sampling method based on the word embedding view, comparing to the spatial distribution or community distribution based sampling.

（case 不具有说服力）

- The first two case studies aren't convincing. I can't agree the authors' claim about Figure 6. "As shown in Figure 6, the obvious differences of human mobility flows in the railway station and residential area can be easily captured". Please add more information about figure 6 and figure 7 to demonstrate the efficiency of the proposed method.

Minor problem

a. In the design of flow wheel, the flow map layout highly depends on the selection of trajectories in the matrix view. What kind of information can we gain through this interactive exploration?（矩阵视图和 wheel 的交互提供了什么信息）
b. In the mobile phone dataset, a trajectory contains several continuous OD flows, which is a reasonable scenario to apply NLP model. Could you please clarify what is the trajectory in the bicycle sharing dataset? In case you consider the OD flows in the same user account as a trajectory, I think the information in the trajectory can't characterize the similarity of OD flows.（语料生成）
c. In Figure 5（c 图编号错误，图片好像放错了），please double check the sub figure labels.
d. In section 5.2, please add sampling rate parameters to table 1（采样率要加上去）. In Figure 9（颜色含义说明不清）, could you explain the meaning of the color and girds?

2.（2.5）

- the writing often lacks clarity. It is sometimes very difficult to interpret.Case studies could not well explain the necessity and usefulness of the proposed methods and the system.

Major problem

（a.主要是图片下方的解释不清）

a. It is difficult to understand each figure because of the lack of explanation of how to interpret the meanings of the figure. Such as,
- The layout of bar chars may cause misunderstanding in a flow wheel. At first, I thought it represents the direction of aggregated OD flows. The meaning of visual mapping should be explained in the captions in the figures (such as Fig.7) to avoid misreading.

- A circular ring（颜色不明显） inside the bar charts in the flow wheel is difficult to find. It should use other color or be explained in the caption of Figure 6.
- What is the meaning of the dark red points in the word embedding view? From the legend shown in Figure 1 (c), I could understand that it represents inter-community flows. But, it should be explained in the caption of the figure or main contents.
- Figure 2 should be explained in more detail in the caption or in Section 3.3. It is very vague. (Section 3.3 explained the tasks in detail but did not explain the figure)
- The explanation in the caption of Figure 8 is too less to understand. It is difficult to find the related explanation in the contents.
（一些术语需要进一步明确定义）

b. The definition of some terms are very ambiguous, so it prevents the accurate understanding of the paper. Each term should be clearly defined in more detail and explained by using examples. Such as,
- correlations of OD flows
- interactions among different flows, interactions of OD flows, semantic interactions of OD flows
- semantic features among flows
- meaningful OD flows
（多目标采样的图探索系统需要引用）

c. Many graph exploring systems using multi-objective sampling have been proposed, they should be cited and differences should be discussed. Such as
- Balancing Systematic and Flexible Exploration of Social Networks
（case 解释不清以及不能证明算法和系统的有效性）

d. Case studies could not well explain the necessity and usefulness of the proposed methods and the system.
（wheel 内部设计）
- I cannot understand the necessity of the flow wheel from the Figure 6. It only shows the very common distributions. It is also difficult to interpret the flows inside the wheel. Authors said that "human mobility flows in the railway station and residential area can be easily captured. In the morning and evening rush hours, the human in the vicinity of the residential area both move from other areas to residential areas. However, the human mobility pattern nearby the railway station is clear in the morning peak, ", But I cannot interpret that way. How to read this?
- There are few case studies showing the effectiveness of the word embeddings (only Figure 8?).

e. Supplemental movie is difficult to understand because there is no caption or narration.

Minor problems:

* In Section 4.3.1, authors said cluster-based edge bundling is integrated into the system, but there is no detail exploration.
* Authors explain that ”The origins of curves are colored in blue while the destinations of curves are colored in red" in Section 4.3.3, but is it right? I cannot find blue and red curves.
* In Figure 5, the labels on the figures are incorrect: (d) -> (c), (f) -> (d).
* In Section 5.5, there is a typo: from 18:00 am to 20:00 am.
* There is some text garbling:
- In section 5.5: can?t identify
- In reference [22]: and ?agatay Demiralp

3.（3）

- The first problem comes from the unclear effects of the NPL-based analysis (Word2Vec technique). In addition, to reduce visual clutter, the samples of the flow movements are displaced several times through the abstraction process, which may distort the original geographic shape of the flows. In this sense, the details of the proposed approach remain to be further investigated.

- The list of references seems to be sufficient, while many of them missing important information such as journal/conference names, pages, etc.
- The list of required tasks together with the system overview is provided in an early stage of the paper exposition.
- The details of the proposed pipeline are adequately described while some notations are used without proper definitions.
- Three case studies are conducted with domain experts, in which the results demonstrate the capability of the proposed approach.
- The paper length exceeds the 9 pages limit by a few lines while this can easily be adjusted through the next revision cycle.

Major problem

- Similarity between flow movements via the NPL-based analysis

(轨迹语义)

As pointed out as the first problem in Section 5.4, the NPL-based analysis of the flow similarity is still limited in the categorization capability of the movement samples because the power of Word2Vec is unsure in the context of this type of analysis. This problem can be further explored. For example, we can try the continuous bag-of-features model instead of the skip-gram model or the hybrid of the two models. The most serious problem may lie in the way of encoding flow movements with NPL terms. The sentence generation depicted in Fig. 3 seems to be reasonable at first glance while it did not work well actually as exhibited in the example of Fig. 8. Investigating several specific failure cases will give helpful hints to alleviate the problem of such improper categorization of flow movement data.

- Possible artifacts in sampling flow movements

  （替换采样点时这个变化通过可视化记录展示）

  Through the proposed abstraction process, the sample points of the flow movements can be replaced several times until the visual clutter will be sufficiently reduced. However, this is likely to distort the original geographical flows and may cause inaccurate visual representations. In practice, how much the original flow movement can be displaced within the proposed framework. Can we set some upper limit for the displacement? It seems that this type of error was recorded through the experiment according to the description, while I prefer to see the results in a more visually plausible fashion.

  (采样算法的最佳情况-最优解)

  Furthermore, we have to find the best combination of available sampling strategies for removing the visual clutter and the combination can change according to the context of the input flow movement data. Even when the system is equipped with an interface for interactively control such combination of sampling strategies, it is quite hard to find the best combination of them without an appropriate guideline. Understanding the visual representations also incurs hard problems. For example, it is a little tough to identify the meaningful differences between the flow patterns with the side-by-side comparison presented in Fig. 6.

c. 视频字幕说明

Minor issues:

Page 3, Left, Section 2.2

"Guo and Zhu and designed a flow selection method to filter out ..."
Guo and Zhu and -> Guo and Zhu?

Page 4, Left, Section 3.3

"..., such as flow intersections, flow importance and community distribution(T.1)."
Insert space ' ' between "distribution" and "(T.1)."

Page 5, Right, Section 4.2.1
"Mathematically, the density at location x is computed by:"
x -> p?

Page 5, Right, Section 4.2.2
"..., if the value of D_{intersec}(p_i) bigger than the D_{intersec}(p_j)."
"bigger than" -> "is bigger than"?
What is 'p_j'? Define it.

Page 7, Left, Section 5.

"..., with the complex data transformation conducted in an off-line preprocessing phase."
How long did it take? Show the statistics on computational times.

Page 7, Right, Section 5.1
"... evening peak and we can?t identify"
we can?t -> we cannot

Page 7, Right, Section 5.1
"It can be found that flow A and the flow B are always retained at different sampling rates."
flow A and the flow B -> flow A and flow B?
Do we have to always try different sampling rates for finding consistent flow movement?

Page 8, Right, Section 5.1

"Compared with the positions in the semantic space, they are divided into two different small clusters semantically."
This is hard to see from the figure.

Page 8, Fig. 8.
It seems that we need to swap the left and right columns in the figure, according to the order of the description in the main text (around the end of Section 5.1).

4. (3.5)

- the failure to demonstrate the utility of the proposed NLP + sampling approach over exiting distribution sampling-based approaches (partly due to unconvincing use cases), as well as incomplete discussion about whether unwanted artifacts/flow distortions may unwittingly be introduced to the reduced OD data.

The Review

Definitions in 4.1.1 need clarification: as I understood it, a single trajectory of a single individual comprises one sentence, all trajectories of a single indivual comprise that individual's document, and the corpus contains as many documents as there are individuals in the data set -- please clarify this by improving and properly defining the index subscripts in formula (1) and its following explanation.

(评审好像理解错了，采样是在投影空间进行的，这点可能需要在论文里说明再清楚一点 | 这个问题没太看懂)
I understood that blue noise sampling is done in high-dimensional embedding space (4.2.1), not in the 2D t-SNE projection, whereas the interactive area-of-interest selection by the user is indeed performed in the projected view (4.3.2). Given the well-documented pitfalls t-SNE projections have in terms of visual analysis of cluster and outlier structures [A], this may introduce unwanted artifacts. Section 5.1 addresses this to an extend, but as also reported in Section 5.3 as feedback from the first expert user this point could receive further consideration (e.g., in Section 5.4).

Flow intersection optimization: this is appartently done in original geographic space. Naive checking for pairwise flow intersection over the entire set of flows would exhibit $O(n^2)$ complexity -- please loose a few clarifying words on scalability wrt. data size, and possible optimizations you included.

Flow importance + and community distribution optimizations: the general idea of applying graph topology measures after OD flows have been extracted from the raw user trajectories has proven to be useful. So while the design choices proposed by the authors for these two specific optimizations are sound by themselves, it begs the question whether this type of imporantce is not captured even better by analysing the raw trajectories (i.e., the actual paths taken by individuals) rather then aggregate flows that do no longer carry information about entire trips. Again, this may be just a matter of slight clarification in the text.

The proposed flow wheel (4.3.3) is very similar to the time lens described in [B], a discussion of the differences to that technique appears to be warranted.

Generally, Section 5 including the use case discussion and expert feedback may benefit from taking a larger portion of the overtall paper, maybe, at the cost of a shortened introduction of the (standard?) graph metrics in Section 4.2 that use a lot of space due to the embedded formulas.

Writing is generally good but some additional proofreading could improve it further (e.g., overabundant use of article 'the'). Caption of Fig.9 appears to be truncated, figure misses a legend.

Note on the supplemental material: video appears to have no audio track which unfortunately diminishes its explanatory effectiveness. Please consider adding either some spoken or subtitle explanations.

Suggested references:

[A] [A] Wattenberg, Martin and Viégas, Fernanda and Johnson, Ian. How to Use t-SNE Effectively, Distill, 2016. doi: 10.23915/distill.00002
[B] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-Based Visualization of Trajectory Attribute Data. IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No. 12, 2012.